14.310x: Data Analysis for Social Scientists
Fundamentals of Probability, Random Variables, Joint Distributions + Collecting Data

Welcome to your second homework assignment! We encourage you to get an early start, particularly if you still feel you need more experience using R. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced. Good luck J!

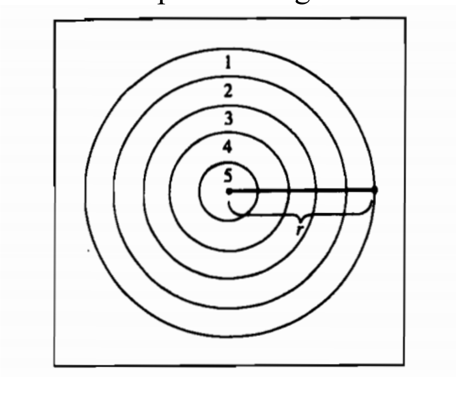
Section 1 – Fundamentals of Probability
Unit 1 – Set Theory and Probability

1. For events A and B in S, which of the following formulas correspond to the probability that either A or B, but not both occur? (Select all that apply)
a. $P(A)+P(B)-P(A\cap B)$
b. $P(A)+P(B)-2*P(A\cap B)$
c. $P(A)+P(B)$
d. $(P(A)- P(A\cap B)) +(P(B)- P(A\cap B))$
e. $P(A\cap B^C)+ P(A^C\cap B)$

Unit 2 – Defining Probability and Examples

2. State whether the following statement is True or False: if $P(A)=1/3$ and $P(B^C)=1/4$, A and B can be disjoint.
a. True
b. False
c. From the information given it is not possible to tell

3. Consider the following example taken from Casella Berger: A game of darts is played by throwing a dart at a board and receiving a score assigned to the region where the dart hits. Figure 1 shows the board and the different possible regions.



Assume that you are a novice player and that a friend suggests that the probability of you scoring *i* points is given by the following formula:

$$P(scoring\ i\ points) = \frac{Area\ of\ region\ i}{Area\ of\ dartboard}$$

Which of the following probability axioms does this problem satisfy? Select all that apply

a. $P(A) \geq 0\ for\ all\ A \subset S$

b. $P(S) = 1$

c. For any sequence of disjoint sets, $A_1, A_2, \ldots, P(U_i A_i) = \Sigma_i P(A_i)$

Unit 3 – Ordered and Unordered Arrangements

4. Using an alphabet of 26 letters, how many unique two-letter initials can be formed if every person has exactly one first name and one surname (last name)?
Note: initials consist of the first letter of the first name, followed by the first letter of the last name. For example, Esther Duflo's initials would be ED.

5. In the game of dominoes, each piece is marked with two numbers and each piece is **unique**. The pieces are symmetrical so that the numbered pair is not ordered: this means that (2,6) = (6,2) and there is only one of such tile. The piece may have identical numbers as well, such as (1,1), (2,2), or (3,3). How many pieces can be formed with different numbers using the numbers 1, 2, …, n?

a. $\frac{2n}{n+1}$

b. $\frac{n(n+1)}{2}$

c. $\frac{n(n-1)}{2}$

d. $n(n+1)$

Unit 4 – Independence and Bayes' Rule

Consider the example you saw in the lecture involving the Zika virus. We will start with the same set-up: A woman lives in a country where only 1 out of 1000 people has the virus. There is a test available that is a **positive result 5%** of the time when the patient **does not** have Zika and a **negative result 1%** of the time when the patient **does** have Zika. Otherwise, it gives correct results. Recall that we computed that th ewoman's chance of having the virus, conditional on a positive test, is less than 1.9%.

In Bayesian parlance, we call the initial, unconditional probability the "prior" and the resulting conditional probability, after updating based on observations, the "posterior."

6. Let the conditional probability we computed (1.9%) serve the role as the new prior. Compute the new probability that she has the virus (new posterior) based on her getting a second positive test. **Please use 1.9% as the prior.**

7. Round your previous answer to the hundredth decimal place. For instance, if your answer is 0.338, you should round to 0.34. How many positive test results would she have to receive in order to be at least 95% sure that she has the virus?

Note: You will need the correct answer from Question 6 in order to obtain the correct response for this question.

a. Two
b. Three
c. Four
d. Five
e. Not possible to infer from the available information

8. In Question 6, we computed the probability of having the Zika virus after a second positive test using the probability of having the Zika virus given a positive test (1.9%). Another way to compute this probability would be to use the unconditional probability of having the Zika virus (1 out of 1000) and treat both the first and second test as independent.

True or False: We would obtain the same probability using either method.
a. True
b. False
c. No possible to infer from the available information

Section 2 – Random Variables, Distributions, and Joint Distribution

For Questions 9 – 15, let $X \sim B(n, p)$, with $n = 8$ and $p = 0.2$.

For this section, please use the R file (hyperlinked in homework on edX) for help with the R code. We highly encourage you to look up the documentation on your own time.

Look at the `rbinom()` documentation to sample draws from this binomial distribution. Use it to generate a vector called "successes" with 1000 draws from this distribution.

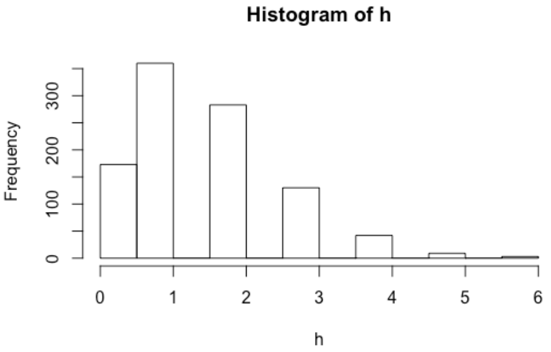
Please note that in R, *n* refers to the sample size and size refers to the number of trials. In contrast, the lectures call the number of trials *n*.

9. Suppose you saved the output of previous step as "successes," which is a numeric vector.

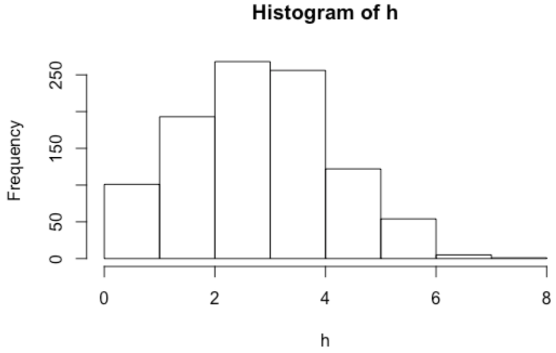
Please fill in the blank to write down a simplest base code to plot a histogram of "successes":

__________(successes)

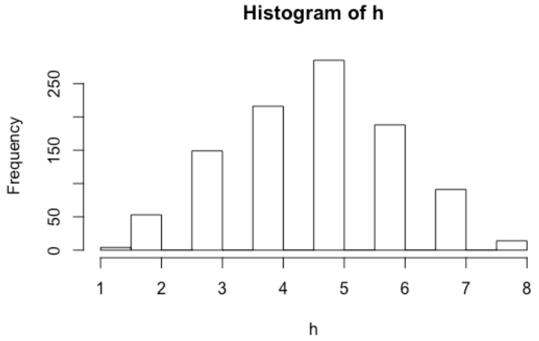
Question 10
Which of the following histograms is closest to the plot that you created?


Histogram of h
Frequency
0
100
200
300
0
1
2
3
4
5
6
h

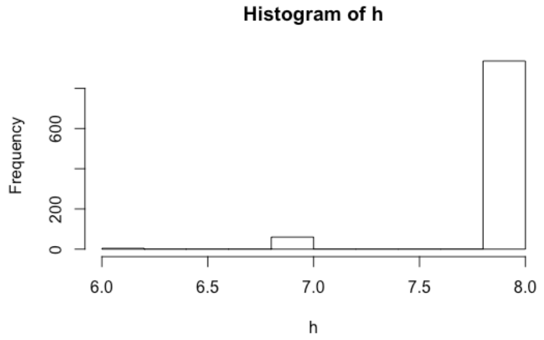
a.


Histogram of h
Frequency
0
50
150
250
0
2
4
6
8
h

b.


Histogram of h
Frequency
0
50
150
250
1
2
3
4
5
6
7
8
h

c.


Histogram of h
Frequency
0
200
600
6.0
6.5
7.0
7.5
8.0
h

d.

11. Try to think about this question without estimating it empirically in R.

Now, suppose $X \sim B(n, p)$, with $n = 8$ and $p = 1$. What is the expected mean and standard deviation of this variable?

12. R has two other built-in functions related to the binomial random variable. One is `dbinom()`, and the other is `pbinom()`. Look up these functions, and use them to answer the questions below.

For parts (a)-(c), suppose you flip an unfair coin, where $p\ (heads) = 0.65$. Round your answer to two decimal places.
(a) What is the probability of getting exactly 7 heads on 10 flips?
(b) What is the probability of getting at most 7 heads on 10 flips?
(c) What is the probability of getting at least 6 heads on 10 flips?

13. If a variable z follows a uniform distribution between 0 and 1, what is the value of the CDF evaluated at any value x if $0 \leq x \leq 1$?

14. Admittedly, the histogram you generated in part 9 could use some work: Firstly, we plotted the frequency counts as opposed to the observed densities in our sample. Recall, that a probability function takes on values between 0 and 1. Second, it is not very pretty to look at.

We found the code below, to plot the densities from the "successes" vector you generated in part 9, however it is filled with blanks. Choose from the drop-down options for the correct code.

```
binom_draws<- as_tibble(data.frame(successes))

estimated_pf <- binom_draws %>%
group_by(_______1_________) %>%
□_______2________(n=n()) %>%
mutate(freq= n/sum(______3_____))
ggplot(estimated_pf, aes(x= successes, y = freq)) +
geom_col() +
ylab("Estimated Density")
```

1. Choose from successes, binom_draws, failures, summarize
2. Choose from mutate, filter, summarize, count
3. Choose from binom_draws, n(), n, successes

15. Instead of plotting the observed density, we could've plotted the analytical densities derived from the formula above. Below is the code to do this. Select from the drop-down options the correct code to fill in the blank.

Please note n=1000 and p=0.2

```
# Create a tibble with x and the analytical probility
densities.
my_binom <- as_tibble(list(x=0:n, prob = dbinom(0:n , n,p)))

# Plot the computed theoretical density.
ggplot(my_binom, aes(x=x, y=prob)) + geom_col() +
ylab("Analytical Density")
```

Now, we are going to use the vector “my_binom” to compute the CDF:

```
calculated_cdf <- my_binom %>%
mutate(cdf = __________)

# Plot the computed cdf
ggplot(calculated_cdf, aes(x=x, y=cdf)) + geom_step() +
ylab("CDF")
```

Select from

a. cummean(x)

b. cummean(successes)

c. cumsum(prob)

d. cummean(prob)

e. cumsum(successes)